FTD-MT-24-1458-71

FOREIGN TECHNOLOGY DIVISION

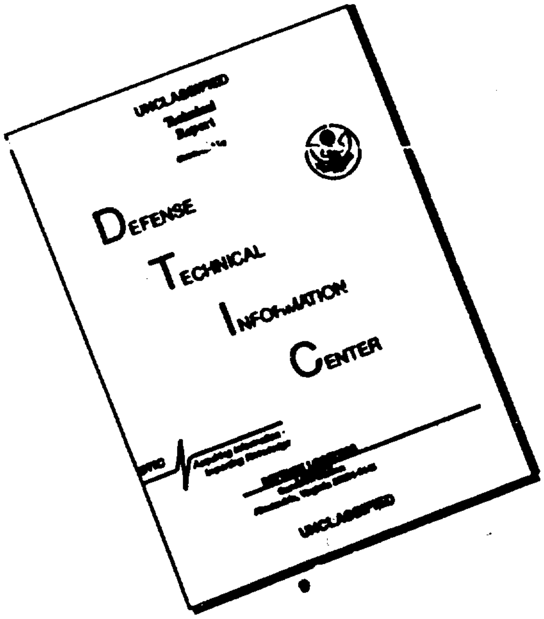
THE PROBLEM OF THE EVALUATION OF RETRIEVAL SYSTEMS, Part I

by

S. Chernyavskiy and D. G. Lakhuti

Approved for public release;
Distribution unlimited.

DISCLAIMER NOTICE



UNCLASSIFIED

Technical Report

DEFENSE TECHNICAL INFORMATION CENTER

UNCLASSIFIED

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Foreign Technology Division Air Force Systems Command U. S. Air Force	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

THE PROBLEM OF THE EVALUATION OF RETRIEVAL SYSTEMS,
PART T

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Translation

5. AUTHOR(S) (First name, middle initial, last name)

Chernyavskiy, S. and Lakhuti, D. G.

6. REPORT DATE 1970	7a. TOTAL NO. OF PAGES 47	7b. NO. OF REFS 8

8a. CONTRACT OR GRANT NO. b. PROJECT NO.	9a. ORIGINATOR'S REPORT NUMBER(S) FTD-MT-24-1458-71
c. DIA Task Nos. T71-05-09 and d. T71-05-13	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AP0116378

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Foreign Technology Division Wright-Patterson AFB, Ohio

13. ABSTRACT

The paper is concerned with a logical analysis of the problem of interpretative validity of the formal evaluation of retrieval systems. The analysis is based on data relating to the SMART system and the Pollack group.

UNCLASSIFIED
Security Classification

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Aerodynamics Center of Gravity Aerodynamic Control						

UNCLASSIFIED

Security Classification

EDITED MACHINE TRANSLATION

THE PROBLEM OF THE EVALUATION OF RETRIEVAL SYSTEMS, Part I

By: S. Chernyavskiy and D. G. Lakhuti

English pages: 41

Source: Nauchno-Tekhnicheskaya Informatsiya. Seriya 2. Informatsionnyye Protsessy i Sistemy (Scientific and Technical Information. Series 2. Information Processes and Systems), No. 1, 1970, pp 24-34.

This document is a SYSTRAN machine aided translation, post-edited for technical accuracy by Charles T. Ostertag.

UR/0447-70-000-001

Approved for public release; distribution unlimited.

THIS TRANSLATION IS A RENDITION OF THE ORIGINAL FOREIGN TEXT WITHOUT ANY ANALYTICAL OR EDITORIAL COMMENT. STATEMENTS OR THEORIES ADVOCATED OR IMPLIED ARE THOSE OF THE SOURCE AND DO NOT NECESSARILY REFLECT THE POSITION OR OPINION OF THE FOREIGN TECHNOLOGY DIVISION.	PREPARED BY: TRANSLATION DIVISION FOREIGN TECHNOLOGY DIVISION WP-AFB, OHIO.

Date 26 Nov 19 71

U. S. BOARD ON GEOGRAPHIC NAMES TRANSLITERATION SYSTEM

Block	Italic	Transliteration	Block	Italic	Transliteration
А а	*А а*	A, a	Р р	*Р р*	R, r
Б б	*Б б*	B, b	С с	*С с*	S, s
В в	*В в*	V, v	Т т	*Т т*	T, t
Г г	*Г г*	G, g	У у	*У у*	U, u
Д д	*Д д*	D, d	Ф ф	*Ф ф*	F, f
Е е	*Е е*	Ye, ye; E, e*	Х х	*Х х*	Kh, kh
Ж ж	*Ж ж*	Zh, zh	Ц ц	*Ц ц*	Ts, ts
З з	*З з*	Z, z	Ч ч	*Ч ч*	Ch, ch
И и	*И и*	I, i	Ш ш	*Ш ш*	Sh, sh
Й й	*Й й*	Y, y	Щ щ	*Щ щ*	Shch, shch
К к	*К к*	K, k	Ъ ъ	*Ъ ъ*	"
Л л	*Л л*	L, l	Ы ы	*Ы ы*	Y, y
М м	*М м*	M, m	Ь ь	*Ь ь*	'
Н н	*Н н*	N, n	Э э	*Э э*	E, e
О о	*О о*	O, o	Ю ю	*Ю ю*	Yu, yu
П п	*П п*	P, p	Я я	*Я я*	Ya, ya

* ye initially, after vowels, and after ъ, ь; e elsewhere. When written as ë in Russian, transliterate as yë or ë. The use of diacritical marks is preferred, but such marks may be omitted when expediency dictates.

FOLLOWING ARE THE CORRESPONDING RUSSIAN AND ENGLISH DESIGNATIONS OF THE TRIGONOMETRIC FUNCTIONS

Russian	English
sin	sin
cos	cos
tg	tan
ctg	cot
sec	sec
cosec	csc
sh	sinh
ch	cosh
th	tanh
cth	coth
sch	sech
csch	csch
arc sin	sin^{-1}
arc cos	cos^{-1}
arc tg	tan^{-1}
arc ctg	cot^{-1}
arc sec	sec^{-1}
arc cosec	csc^{-1}
arc sh	sinh^{-1}
arc ch	cosh^{-1}
arc th	tanh^{-1}
arc cth	coth^{-1}
arc sch	sech^{-1}
arc csch	csch^{-1}
rot	curl
lg	log

THE PROBLEM OF THE EVALUATION OF RETRIEVAL SYSTEMS, PART I

S. Chernyavskiy and D. G. Lakhuti

INTRODUCTION

This work is devoted to the problem of the evaluation of retrieval systems. Generally speaking, it would be possible to differentiate two forms of evaluations of retrieval systems, which it is convenient to call respectively external and internal. External evaluation is always based on the comparison of the results of the operation of a retrieval system with a certain "ideal" result, and therefore it always uses a concept of relevancy. Internal evaluation should be based on the properties of the retrieval system such as, for example, complexity, the degree of nearness to human logic or natural language, etc. - and not use a concept of relevancy. Reasonable internal evaluations of retrieval systems are unknown to us and in the present work we are limited to an examination of external values. It seems that in the problem of external evaluations it is possible to set aside three fundamental aspects. The first of them is connected with the concept of relevancy and, especially, with the question of the means of determination of relevant distribution. The second aspect is connected with the question of the high quality of those measures which are selected for the evaluation of retrieval systems. The third aspect is connected with the question of the statistical reliability of the computable values of evaluation or, in other words - with the question of the representativeness of those sampling groups of requests and documents on which these values are calculated.

As far as we know, up to now only the first of these aspects has been the subject of systematic special investigations (specifically in the works of the Cleveland group in the USA). Our article is devoted to the second aspect of the problem of values, which deserves the most fixed attention.

We consider that the fact that it is necessary to be able to evaluate retrieval systems both by themselves and in comparison with each other does not require proof. Today many very diverse measures are known which have been proposed for the evaluation of retrieval systems. They are frequently incomparable, they frequently contradict each other, and their authors, it turns out, speak about the deficiencies of others and the advantages of their measures, but it is significant that they never receive an answer: no one accepts the call and controversies are not obtained. The impression is created that there is simply no base for controversy: there is no clear understanding of which requirements should be satisfied by "good" measures and which principles should be followed during their comparison and sampling. We fear that without such an understanding the organizing of new measures and values as well as the utilization of the old, to a considerable degree risks becoming depreciated.

Without claiming to have a solution to this problem, the present article at least places it in evident view and indicates some significant, from our point of view, aspects of it. We will illustrate our analyses with material taken from the works of SMART[1] [1, 2] and Pollock [3], which unquestionably belong to the ranks of most solid works in the area which interests us. Subsequently we hope to analyze the other works, and in the first place - the interesting work by Cooper [8], from the same point of view.

§ 1.

Here those concepts are introduced which it is necessary to introduce prior to the beginning of an investigation. The others will be introduced when it is necessary.

[1] For simplicity we will use the abbreviated name of the well-known American project SMART as the proper name of the author of the publications connected with this project.

The functions of nearness and distribution. At the basis of our examinations lies two concepts, closely coupled with retrieval systems - the functions of nearness and distribution.

The function of nearness compares to every inquiry - document pair some number (or another object of an abstract nature) characterizing the relevancy of the given document based on the given inquiry **from the point of view of the retrieval system.** If we, having fixed the inquiry, arrange the group of documents in decreasing values of nearness assigned by the function of nearness, then we will obtain that which we call the (relative) distribution of documents on the given inquiry.

The functions of nearness can be of two types depending on the number of possible values assigned by them to the documents. If the number of values of the function of nearness is great, so that every document of the group can be assigned its own, different from all others, value of nearness, then such a function we will call a **function of the SMART type.** If the number of possible values of the function of nearness is small, and therefore in a normal case substantially less than the number of documents in the group,[1] so that one and the same value of nearness is assigned to many documents of the group, then we will call such a function a **function of the Taub type.** It is clear that the SMART functions of nearness give distributions which are a complete ordering of the group of documents, while the distributions generated by the Taub functions give only a partial ordering. In accordance with this we will designate as SMART and Taub not only the functions of nearness, but also the distributions generated by them, being distracted from the fact that, let us say, the SMART function can, abstractly speaking, generate Taub distribution, and vice versa.

The separation of functions and distributions into SMART and Taub which was introduced by us is not strict and, generally speaking,

[1] We have in mind normal cases of retrospective retrieval, and not the directional distribution of information, when the groups can be very small.

indefinite intermediate cases are possible. Nevertheless we consider it natural and useful because, in the first place, all the known functions of nearness and the distributions generated by them in real situations are either evidently SMART or evidently Taub, and since, secondly, in the terms of this discrimination the intrinsic properties of the values of retrieval systems are described.

Having a certain function of nearness irrespective to SMART or Taub, we can number (all or some) its values in decreasing order. We will call such numbers **ranks**. If all the possible values of a function are numbered, then we call the ranks **absolute**, if only the values of nearness actually assigned to any documents (from the given inquiry) are numbered, then we call the ranks **relative**. The distribution of documents by absolute ranks is called **absolute distribution**, and distribution by relative ranks coincides with a previous introduction by relative distribution. Thus by definition in the absolute distribution of documents based on the given inquiry empty ranks are possible, which is impossible in relative. Let us note that in the language of ranks Taub distributions differ from SMART by the fact that in their ranks there can be more than one document.

Let us clarify what was said by examples. The function of nearness does not always take numerical values; Taub functions of nearness - if we do not resort to artificial methods - never take numerical values. Thus in any single-term set the function of nearness takes in the case of the criterion of distribution "for complete entrance" two abstract values. Corresponding to these are two ranks - the issued and nonissued documents. The first of these ranks can be empty, whereas the second in all actual cases is unempty. In the system "empty-not empty-2" [Pusto-Nepusto-2] the function of nearness takes three nonnumerical values which correspond to three ranks - "yes," "maybe" and "no." Nonissued documents correspond to the last of these ranks. Any of the first two ranks can be empty, the latter in any real case is unempty.

Let us consider the following example. Assume for a certain inquiry one of the two systems of the type "empty-not empty-2" assigns to any two documents the value "yes," and to all the others - the value "no," and the second system assigns to the same two documents the value "maybe," and to all the others - the value "no." In this case we have two different absolute distributions with three ranks, whereupon in the first of these distributions the second rank is empty, and in the second - the first rank, and two coinciding relative distributions with two ranks, which have in the first (relative) rank two documents, and in the second - all the others.

A SMART type typical function of nearness is the so-called "function cosine" which is used extensively in the works of SMART. It is determined in the following manner:

$$\cos(q, d) = \frac{\sum_{1}^{n} q_i d_i}{\left(\sum_{1}^{n} (d_i)^2 \cdot \sum_{1}^{n} (q_i)^2\right)^{1/2}},$$

where q and d are n-dimensional vectors in the space of descriptors which represent inquiry q and document d, and q_i and d_i - respectively their i-coordinates, which take the values 1 or 0 depending on whether or not the i descriptor enters into the appropriate descriptor pattern (Smart himself also uses, apart from one, other weights which are different from zero). This function can take both rational and irrational values between 0 and 1. In practice the numerical values of the function of nearness are always measured with a certain final accuracy. If, for example, we consider the values of nearness to within the sixth decimal point, then 10^6 absolute ranks are obtained, which, strictly speaking, also are in mind when we are speaking of the absolute distributions which correspond to the function in question.

We will return to the fundamental presentation. During the functioning of a retrieval system the distribution of documents is used for the organization of distribution of documents based on the inquiry to man. This can be done in two ways.

In the first method man is offered the entire group of documents, ordered with the help of distribution. In this case man - the user who assigned the inquiry, or the clerk operating the system - looks over the documents in the ascending order of ranks and in some way or other determines the moment for the curtailment of scanning. The documents scanned up to this moment are naturally considered as the formal distribution of the system. Thus in the described circuit of functioning the system does not determine the formal distribution for the inquiry unambiguously. In such cases we will say that the system works in an **incomplete** mode.

In the second method of the utilization of distribution in the system a threshold which separates the issued part of distribution is assigned. Depending on the distribution utilized the threshold can be assigned in terms, let us say, of absolute or relative ranks, but in any case without the utilization of the concept of relevancy. In such a circuit of functioning the system itself determines the formal distribution completely, and in this case we are speaking about the **complete** mode of operation.

Let us note that in the case of incomplete mode it is significant, it goes without saying, not that a certain operation is executed not by machine, but by man, but that in this case he uses any facilities available only to him; in the case of a retrieval system such a facility is the relationship to relevancy. Therefore any criterion of distribution which uses a concept of relevancy - perhaps along with other purely formal facilities - makes the mode incomplete.

We will consider that a retrieval system consists of a language and a system of indexing, and also the function of nearness; during operation in the complete mode a certain threshold which determines formal distribution;[1] during operation in the incomplete mode we can speak about formal distribution only after we are assigned the criterion in one way or another. By using it man determines the moment of curtailment of scanning. Perhaps the aforesaid should not be considered as the definition of the concept "retrieval system,"

[1] The function of nearness together with threshold make up that which in the other operations (see for example [4, 5]) we call the "rules of comparison" or the "criterion of distribution."

if we require of the definition that it makes it possible to distinguish the cases when we are dealing with one and the same retrieval system from those cases when we are dealing with different ones. We will not begin now to occupy ourselves with the appropriate specifications, although this leads to certain terminological nonstrictness. Thus, for instance, we will not make distinctions between the expression the "system which works in a complete (incomplete) mode" and the expression "complete (incomplete) system," although in the second method of expression we are speaking about two different systems where in the first method we spoke about one. Within the framework of the present analysis this should not lead to misunderstandings.

The classification of systems. Taking into account everything expounded it is possible to speak about retrieval systems of the Taub or SMART type depending on the type of the function of nearness utilized in the system; about absolute and relative retrieval systems - depending on the type of distributions (ranks) utilized in the system; and finally about complete or incomplete retrieval systems - depending on the presence or absence of a threshold of distribution.

The question of the practical realizability of systems of the types enumerated deserves at least a brief consideration.

The majority - if not all - of systems operating now are absolute and complete. A system with a mixed criterion of distribution, in which absolute ranks arc used along with the concept of relevancy, can be absolute and incomplete. Such a system is possible to imagine, but we know nothing about operational. systems of such a type or of any under development. Further, although it is possible to imagine a relative system which works in a complete mode, the naturalness of such a situation is completely doubtful, since in this case the threshold would have to be formulated in the terms of the number of ranks issued, i.e., to issue a fixed number of ranks for any inquiry. Apparently an alternative could be only a threshold, computable based on the inquiry, which today is unreal. As concerns incomplete systems, then in practice Taub systems, incomplete in the true sense of this word, are not encountered. This is explained by the fact that among the sparse ranks in the distributions of these systems there is always one - the last, which knowingly contains all the documents which are unnecessary for the inquiry and therefore is extremely great. One cannot issue this rank, and there is no sense to issue any of the comparatively small number of others - therefore as a rule they are all issued, and the partial order of documents assigned by them facilitates the examination. Thus, in these systems actually there is always a threshold - before the last rank.

SMART systems remain. Although in contemporary literature, especially in the works of Smart ([2], p. 27) and Pollock ([3], p. 393), the opinion is stated that SMART functions of nearness are most natural for automatic retrieval systems, and although systems of the SMART type are the object of a significant number of theoretical analyses and experimental investigations, their industrial realizability in retrospective retrieval systems with more or less considerable groups causes serious doubts. The fact is that in SMART systems the adding of distant ranks requires as much time as the adding of close ones, whereas in Taub systems the reference of a document into the last most numerous rank, containing a suppressing number of documents, usually requires considerably less time than its reference into one of the first, containing only an insignificant part of the group. The impression is created that on sufficiently large arrays of documents SMART systems will substantially lose out to Taub from the point of view of retrieval time, and also because of the knowingly unnecessary operation spent on the discrimination of ranks of documents knowingly not scanned by man.

It is possible, however, to assume that incomplete SMART systems can find use in the distribution systems of current information with their small arrays of documents or in systems of two-stage retrieval, when during the first stage the overwhelming majority of array is detached by Taub methods, and in the remaining comparatively small subarray the SMART function of nearness completely orders the document, representing to man their complete or partial scan.

Evaluation. In the present work we differentiate two types of formal values: evaluation-scale and evaluation-description. Every formal evaluation is a certain effectively computable operator which to every evaluated object compares a certain other object called the **value** of evaluation. From the evaluation-description it is required that its values make it possible to judge sufficiently fully the important properties of the evaluated objects, for example, to forecast their behavior under any conditions, and in this case we call the evaluation-description **effective**. From the evaluation-scale it is required that its values would order the set of evaluated objects, without entering into contradiction with our existing meaningful representations about the comparative advantages of these objects, and in this case we call the evaluation-scale **sensible**. One ought to keep in mind that one and the same formal evaluation can be examined and used both as an evaluation-scale and as an evaluation-description.

When there are no foundations for misunderstandings we will speak simply about (formal) evaluations, adding the definitions "scale" and "description" only in the case of real need. Evaluation-description will be the subject of our examination only in the third part; now we will be occupied with evaluation-scale.

Evaluation-scale. From a sensible evaluation-scale for retrieval systems we require the following. Primarily the values of an evaluation should form at least a partially ordered set and therefore induce at least partial order in the set of retrieval systems. Further, this partial order should not contradict our meaningful representations about the comparative advantage of the various systems in those cases when we have such representations. Finally it is also possible to require that in those cases when we have meaningful representations about the comparative advantage of two retrieval systems, the values of evaluation assigned to these systems are also congruent. Meaningful representations about the comparative advantages of systems we will also call **meaningful evaluation**. Thus it can be said that a sensible formal evaluation should not contradict meaningful.

Any evaluation is oriented on some properties of the evaluated systems. In the present work basically the semantic properties of retrieval systems are examined, and such important properties of them as, let us say, the cost of operation, the laborinput for development, etc., are not examined.

Here it is appropriate to make a certain digression.

In the first place it follows from the aforesaid that if the high quality of a formal evaluation-scale is considered to be coinciding with its **soundness**, formally determined above, then this high quality substantially depends on the meaningful point of view for the comparative advantages of the evaluated retrieval systems. This point of view is determined by that problem for which we wish to use the retrieval system, and by those conditions in which we use it. Hence it obviously follows that with a change in conditions and the problem being solved the point of view can be changed, and with it the formal evaluation used. In spite of the apparent obviousness of this confirmation, none of the investigators known to us who are engaged in the problem of evaluating retrieval systems are making sufficient conclusions from this. This is manifested in the fact that none of them considers it necessary to actually connect the formal evaluations introduced by them with any certain conditions of their application. At the best some of them, Smart and Pollock for example, are limited to a reference to the possible dependence of formal evaluations on the meaningful points of view. We will show this in concrete examples in the course of the following presentation.

On the other hand, it should be noted that although all the authors who are writing about evaluations are speaking about them as about the evaluations of namely **retrieval systems**, in reality they frequently speak about them or they use them for the evaluation not of retrieval systems by themselves, but for the evaluation of retrieval systems under certain concrete conditions of functioning, i.e., which for us is convenient to call **retrieval service**. In the present work we do not have the possibilities to engage in particular on the question of the demarcation of these two approaches and therefore the examination is conducted in a general form, encompassing both of them.

We will now return to tne fundamental presentation.

If we are limited, as this is done in the present work, to the semantic or its related properties of systems, then it is sufficiently natural to construct the evaluation of system on the basis of the individual distributions which are generated by system, or their values. The evaluation-scale for distributions is determined in the same manner as the evaluation-scale for retrieval systems, and the concept of soundness introduced above extends to it.

Any distribution is determined, apart from the retrieval system, also by concrete inquiry and the array of documents. Therefore for obtaining the evaluation strictly of the system it is necessary in one way or another to eliminate the arrays of inquiries and documents (if this actually is possible). This elimination in practice always takes the form of some averaging. In this case it is natural to consider especially that if the evaluation of system is obtained by the averaging of the evaluations of distributions and if all the evaluations being averaged are equal to each other, then the evaluation of system is equal to this common evaluation of distributions.

§ 2.

In this section, speaking about evaluations, we will always have evaluation-scale in mind.

Tne normalized evaluations of Smart. Let us consider the pair of evaluations for distributions which were introduced by Smart (comp. [2, 3]) under the name **normalized completeness** and **normalized accuracy**.

Normalized completeness R_N is determined by the formula

$$R_N=\frac{1}{N}\sum_{i=1}^{N}\frac{n_i}{n}, \qquad (1)$$

where N - number of documents in the array; n - number of necessary (relevant) documents in the array; n_i - number of necessary documents up to i-th (relative) rank inclusively.

From formula (1) it is evident that the number of ranks is assumed equal to the number of documents, which corresponds to the definition of relative Smart distribution.

The normalized accuracy P_N is determined by the formula

$$P_N=\frac{1}{N}\sum_{i=1}^{N}\frac{n_i}{i}, \tag{2}$$

where N and n_i - previous, and i - relative rank.

Taking into account that in formula (1) it is possible to carry out 1/n beyond the summation sign and that every relevant document adds into the common sum as many units as ranks after it plus 1, formula (1) can be rewritten as:

$$R_N=\frac{1}{n}\cdot\frac{1}{N}\sum_{i=1}^{N}m_i\,(N-i+1), \tag{3}$$

where

$$m_i = \left\{ \begin{array}{l} 0, \text{ if i document is not relevant, and} \\ 1, \text{ if i document is relevant.} \end{array} \right\}$$

If now we consider a document of rank i with a point with mass m_i arranged on the axis at a distance i from the origin of reading, then formula (3) shows that $R_N N$ is taken with the reverse sign of the coordinate (relative to point N + 1) of the centroid of the system of these objects. Hence it follows that R_N is invariant relative to any redistribution of documents by ranks which does not change the position of the centroid (in our case - not changing the sum of the ranks of the necessary documents).

We will turn now to evaluation (2). Let us assume that during a certain distribution at i place there is an unnecessary, and at (i + 1) a necessary document. If we exchange their position, then in formula (2) only i and (i + 1) terms are changed. Prior to their transposition the sum was equivalent to

$$\frac{n_i}{i}+\frac{n_i+1}{i+1},$$

after transposition it will equal

$$\frac{n_i}{i} + \frac{n_i+1}{i+1} = \frac{n_i}{i} + \frac{n_i+1}{i+1} + \frac{1}{i}.$$

Hence it is clear that during the transposition of two adjacent documents, of which one is necessary and the other unnecessary, P_N is changed by a value, inversely proportional to the place of transposition. From this, in particular, it follows that the extending of the system of necessary documents, without changing the position of the centroid of distribution, improves P_N without changing R_N.

Now let us present two distributions - D1 and D2 [Д1, Д2] such that D2 is obtained from D1 by the extension described above without changing the position of the centroid of distribution. According to what was proven above

$$R_N(Д1) = R_N(Д2),$$

and

$$P_N(Д1) < P_N(Д2).$$

For illustration let us assume, for example, that D1 has the form

$$- - + + - - + + - - \ (Д1),$$

where the sign "+" designates necessary and the sign "-" unnecessary document, and the ranks are numbered from left to right. Further assume D2 is obtained by the symmetrical extension of the system of necessary documents on one rank

$$- + + - - - - + + - \ (Д2).$$

What can be said about the comparative advantage of the distributions in question, on the strength of the values of evaluations obtained by them? It should be noted that Smart introduces - as far as it is possible to judge - the evaluations R_N and P_N not for their separate use, but (similar to ordinary completeness and accuracy) as a pair, which should be used for the comparison of distributions and systems as a unit (comp. [1], p. 213, 215). Such a combined utilization of several particular evaluations for the construction on their basis of any joint evaluation can be accomplished by different means. Specifically Smart proposes two such methods ([1], p. 216) which we do not consider here. However, for any joint evaluation it is natural to consider that if for one object the values of all particular evaluations are not lower, and the value of at least one particular evaluation is higher than for another object, then the value of the joint evaluation for the first object is higher than for the second (i.e., joint evaluation monotonically increases together with every component evaluation). Hence, in the example in question one ought to consider that D2 acquires a higher joint evaluation than D1.

In order to judge the soundess of the evaluation (R_N, P_N) we should now compare the conclusion obtained with the meaningful representation about the comparative advantage of distributions D1 and D2. Which of these distributions is better from a meaningful point of view? This depends, it goes without saying, on the point of view. For certainty let us assume that user formulates his point of view, indicating the number or the fraction of necessary documents which he wishes to obtain or which he is prepared not to obtain. In this case the point of view of user determines in the terms of number or fraction of necessary documents a certain optimum threshold in distribution. The documents located to the left of this optimum threshold should be issued, and the documents located to the right - no, since the user, by hypothesis, will not evaluate the increases in completeness over the border indicated him, but will be dissatisfied by the additional noise.

Let us assume that the conversion which converts D1 into D2 affects all the necessary documents. Then, if the optimum threshold, determined by the point of view of the user, lies to the left of the centroid (for instance, if in our illustration it is required to give out two necessary documents), then after conversion it will be moved still more to the left, thereby decreasing noise, so that D2 is meaningful better, than D1, If the optimum threshold lies to the right of the centroid (for instance, if in our illustration it is required to give out three necessary documents), then after conversion it will be moved still more to the right, thereby increasing noise, so that in this case D2 is meaningful worse than D1. Thus in the first situation (the optimum threshold to the left of the centroid) the evaluation (R_N, P_N) on distributions D1, D2 is sensible, but in the second situation (the optimum threshold more to the right of the centroid) - it is not.

It follows from this example that it is not possible to speak about the soundness of any joint evaluations of retrieval systems based, as on components, on Smart evaluations of R_N and P_N without making more precise the point of view which determines the meaningful evaluation of these systems. This means that this type of joint evaluation should not be introduced without refining their "area of soundness," i.e., without describing one way or another those points of view, at which the input evaluations can be used meaningfully. At the same time Smart, although he mentions the possibility of various points of view on the meaningful evaluation of retrieval systems (comp. [2], p. 11, 12), makes no conclusions - neither practical nor theoretical - from this.

In the works [1] and [2] Smart, apart from (R_N, P_N) introduces two additional pairs of analogous evaluations which he considers to be more convenient in computations. The example with distributions D1 and D2 relates completely to both these pairs with an insignificant change in the proofs. Therefore we will not dwell on these evaluations separately.

Universal evaluation-scale. A natural question appears: in general is a universal evaluation possible, universal in that sense that its area of soundness includes all the meaningful points of view?

The answer to this question - and moreover negative - can be considered as our example with D1 and D2. Nevertheless,for reasons which will be presented at the end of part 5, we will consider this question from another point of view.

First of all let us refine the concepts used. We are speaking about (formal) evaluations of retrieval systems and the distributions generated by them, and also about (meaningful) points of view for the comparative advantages of these retrieval systems and distributions or, in other words, about the meaningful evaluations of these systems and distributions. Any point of view, or a meaningful evaluation, orders (generally speaking partially) the set of retrieval systems or distributions and determines soundness or unsoundness of the corresponding formal evaluation: and namely, an evaluation is sensible in that and only in that case when the order assigned by it coincides, or at least does not contradict the order determined by the point of view (the meaningful evaluation).

Generally speaking the concept of the point of view of meaningful evaluation is sufficiently indefinite and can be refined by various means. Within the limits of the given work we will consider that the content of this concept is exhausted by the ordering of systems or distributions assigned by it.

We will consider that any point of view for retrieval systems includes one or another point of view for distributions. In accordance with this in the present work only those evaluations of retrieval systems are examined which by some means are constructed from some characteristics - perhaps evaluation-distributions. Let us note that this type includes all evaluations of retrieval systems known to us. The question of the possibility of other types of evaluations we do not consider.

The points of view for retrieval systems can be connected with the points of view included in them for distributions by various means. However, it is natural to assume that in all reasonable cases a monotonicity of the meaningful evaluation of system will take place in its own way according to the meaningful evaluation of distributions. This should be understood thusly: assume we are meaningfully comparing two retrieval systems for one and the same array of inquiries and documents. Then we obtain two arrays of distributions, between which a natural one-to-one conformity exists. Now if all the distributions of one system are appraised meaningfully no lower than the corresponding distributions of the other, then the first system cannot be evaluated meaningfully lower than the second; and if at this any distributions of the first sytem are meaningfully appraised higher than the corresponding distributions of the second, then the first system will be evaluated meaningfully higher than the second. It is clear that from the determination of the soundness of the evaluations of systems and distributions it follows that for sensible values an analogous monotony is preserved. Further it can happen, that we will want to compare two systems both for various arrays of documents and inquiries (we will speak about the sense of this below). Then the requirement of monotonicity changes form somewhat. Namely in this case it is natural to assume that if the minimum of the meaningful evaluation of the distributions of the first system is higher than the maximum of the meaningful evaluation of the distributions of the second system, then the first system will obtain a meaningful higher evaluation than the second. And in this case from the determination of the soundness of formal values it follows that the corresponding monotonicity will take place even for the sensible formal values of systems. It is clear that the formulated requirements of monotonicity are not independent: when it makes sense and the first requirement is carried out, the second is also executed.

Now let us consider the degenerate special case, in which all the distributions of the first system are equal to each other, just as also all the distributions of the second system are equal to each other. Since we assume that in such cases the evaluation of the system coincides with the evaluation of the corresponding distribution, then it is not difficult to see that relative to the given point of view, no evaluation can be sensible for retrieval systems built with the utilization of an unsound - relative to the same point of the view - evaluation of distributions.

Before we continue our examination, let us make two remarks.

In the first place what was said above about the connection of meaningful points of view and formal evaluations is in essence a strict substantiation of that application which we gave to our example with distributions D1 and D2.

In the second place, we always proceed from the assumption that the soundness of an evaluation is determined by the point of view, so evaluation is recognized as sensible when and only when it is sensible from this viewpoint for any possible distributions. However, it would be possible to consider that soundness is determined not only by the point of view, but also by the class of permissible distributions. In this case one should speak about the soundness of a formal evaluation not relative to one or another point of view, but relative to the situation which, in addition to the point of view, includes distribution. Then the area of soundness would be a certain class of situations which in practice apparently should always be given one or several sets of threes of objects: point of view, the array of inquiries, the array of documents. In the present work we will be limited to the first approach.

We will return to the question of the possibility of universal evaluation.

Every point of view in a natural way determines a certain group of transformations of distributions, relative to which the meaningful evaluation, which is the expression of this point of view, should be invariant. This should be understood so that if two distributions are obtained one from the other with the help of one of the conversions of group, then the values of evaluation assigned by these distributions are equal. In this case, however, it is not assumed that any conversion of a group is applicable to any distribution. For brevity we will say that relative to the conversions of a group point of view itself is invariant. Then it is clear that every sensible formal evaluation relative to this point of view also should be invariant relative to the same transformation group as also the point of view. Thus, for instance, every reasonable point of view on distributions should be, apparently, invariant relative to the arbitrary transpositions of documents of equal relevancy. Therefore, relative to this transformation group any sensible formal evaluation of distributions should also be invariant.

Let us consider the conversion which we call the similarity transformation. Assume we have a certain retrieval system, a certain inquiry, and a certain array of documents. Then we have a certain distribution D3. Now we replace every document of our array by m of the same (or almost the same) documents so that any two documents, generated by one initial one, would on the scale of absolute ranks always be considerably closer to each other than any two documents generated by different original documents. We will say that the distribution D4 of this new array on the same inquiry in the same retrieval system has been obtained by the similarity transformation from distribution D3. Thus, the similarity transformations determined form a group, among the elements of which there are those which are not applicable to any distribution. However, in the given context this is unessential for us.

Reasonable points of view exist which are invariant relative to the determined similarity transformation. We contend with such a point of view every time, when we wish to evaluate any retrieval system by itself, as far as possible independent of the dimension of the array with which it works.

On the other hand, the reasonable points of view exist which are not invariant relative to such a similarity transformation. We contend with such points of view when we wish to use the formal evaluation in order to consider a certain retrieval system in connection with various arrays of documents, for example, in order to select for the given system the permissible field of application, i.e., actually for the evaluation of the retrieval service. For example, assume we use a Smart type incomplete retrieval system so that for every inquiry we find at least one (any) relevant document if it is in the array.

In this case we are not interested in the presence of other relevant documents, but we wish that the number of unnecessary documents preceding the first rele ant one be as small as possible. This point of view we will conditionally call the **"point of view of patent newness."** With such a formulation we deal still from the point of view for a **system** and it is natural to consider it invariant relative to the similarity transformation. If to this formulation we supplement the requirement that the number of unnecessary documents, preceding the first relevant one does not exceed a certain fixed number, which determines the boundaries of the practical applicability of the system, then we obtain the point of view of "patent newness" already for the retrieval **service.** It is clear meaningful evaluation which is determined by this last point of view cannot be invariant relative to the similarity transformation because here we would be deprived of a single criterion which would permit us to determine, let us say, the maximum dimension of the array of documents, at which it is still possible to use a given system in practice.

It follows from the aforesaid that a formal evaluation, the field of soundness of which would encompass both points of view described above - the point of flew for the system and the point of view for the service cannot exist, and in this sense one ought to answer negatively to the question of the possibility of universal evaluation. It is possible, however, to narrow down the question and to be limited only to those situations, when we wish to evaluate retrieval systems by themselves, as far as possible independent of the array of documents and inquiries, i.e., to consider the question of the possibility of the universal evaluation of retrieval **systems** (in contrast to **services**).

In order to answer the question of universal evaluation which is recognized, let us introduce into examination one additional point of view which we conditionally call the **"point of view of patent purity."** With this point of view the user wishes to obtain all the relevant documents, and its meaningful evaluation of the results of the activity of the retrieval system endures a significant jump during the transition from zero losses to nonzero. If now we compare the twc "patent" points of view for the system, then it is easy to see that the first is invariant relative to any conversion, not affecting the rank of the first relevant document and the entire field of ranks preceding it, while the second is invariant to any conversion in the field of ranks, fcllowing after the rank of the last relevant document, but is not invariant, let us say, to the movement of the last relevant document by one rank to the right. Thus in this sense also a universal formal evaluation is impossible.[1]

We do not eliminate the fact that the further development of the problems of the evaluation cf retrieval systems leads to the examination of new types of values, to other definitions of soundness or to other classifications of the points of view and situations.[2] In this case it can happen that the conclusions just made by us will be unimportant. But one thing is indisputable: it is senseless to introduce or to use any formal evaluations, without indicating or without investigating simultaneously the area and the conditions of their intelligent use; the absence of such indications in contemporary works dedicated to retrieval **systems** and services is one of the fundamental reasons which block the meaningful utilization of the proposed values.

[1] At the same time apparently evaluations are possible in which the area of soundness is empty. As a limiting case let us give the evaluation proposed by D. Yu. Teplov in his doctoral dissertation (the dissertation was defended on 16.05.69 at the Leningrad Institute of Culture imeni N. K. Krupskoy). The evaluation has the form u'/u. Here u' - the number the "unexpectedly valuable" documents issued by the system, and u is equal to $\sum_{1}^{n} R - R_u$. where R - "pertinent distribution," R_u - "irrelevant distribution," and n - the "number of courses in the strategy of retrieval."

Unfortunately this determination (Vol. II, p. 470 of the dissertation) has much that is unclear. Especially unclear is whether R_u is located under the summation sign or not. However, in any event this evaluation possesses many surprising features, among which let us note only one. The graph of the dependence of evaluation u'/u on on Ru (respectively from Σ R_u i.e., on the absolute value of noise, takes the following form (Fig. 1).

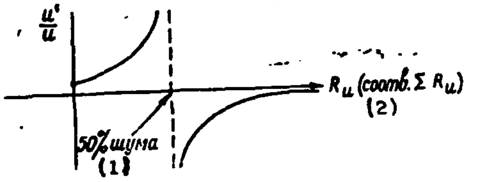


Fig. 1.
KEY: (1) noise; (2) correspondingly.

Thus the evaluation encourages the distribution of noise. We could not fabricate a reasonable point of view which would enter the area of soundness of this evaluation.

[2] One of the interesting trends of research it seems to us is, for example, the transition to parametric evaluations which would be able to consider the change in the point of view by the change in the parameters included in them. This idea, true in a quite limited sense, was formulated by Pollock ([3] p. 391). We will speak about this below.

Today as formal values ordinary completeness and accuracy are used most often. In this case the authors, giving the values of these evaluations which were obtained for concrete cases, never pose the question of whether or not their utilization in the given context is appropriate in general. Although the precise determination of the area of soundness for completeness and accuracy, and also for the joint values constructed from them is difficult today, it is clear in any case, that both these evaluations - invariant relative to the similarity transformation - cannot be used by themselves, for example, in those of the situations described above which do not allow such an invariance.

The problem of the connection of formal evaluation and its field of soundness can be examined even in a reverse setting: namely, is it possible from the assigned point of view to seek the corresponding evaluation. Not having the effective apparatus for the solution of such problems, we will be limited to an example.

Let us stand on the point of view of "patent newness" determined above and from this viewpoint we wish to evaluate a certain complete absolute Taub retrieval system by itself, as far as possible independent of the arrays of documents. Then, for example, the following evaluation of distributions is sensible:

$$t \cdot sgn\,(p) - k\,[1 - sgn\,(p)];$$

where t - accuracy; p - completeness; k - penalty for zero completeness;

$$Sgn\,(x) = \begin{cases} 1 \text{ при } x > 0 \\ 0 \text{ при } x = 0 \\ -1 \text{ при } x < 0 \end{cases}$$ при = at

§ 3.

Here we investigate new evaluations which we will examine not only as scales, but also as descriptions. As it was shown above, the reasonable use of the evaluation-scale is determined by its soundness. For evaluation-description an analogous role is played by the concept of effectiveness: we call the evaluation-description effective for the given retrieval system, if the value of this evaluation, calculated for the given system, actually makes it possible to say something useful about the behavior of the evaluated system. The concept of effectiveness is considerable less definite than the concept of soundness, however, in concrete cases its utilization does not cause inconveniences.

Let us consider the evaluation, also introduced by Smart (see [2], p. 11-14). In contrast to previous ones the values of this new evaluation are not numbers, but graphs. This evaluation of retrieval systems, to a greater degree than the previous ones, can be considered as an evaluation-description, although it can also be used, and it is actually used by Smart, for the comparison of retrieval systems, i.e., as an evaluation-scale. Smart examines basically not the graphs of individual distributions, but the graphs averaged for the inquiries, which are the values of evaluation for the retrieval systems. Correspondingly in this section we will give primary attention to the question of the dependence of soundness and effectiveness of evaluation of a system on the means of its construction from the evaluation of distributions.

The graph of Smart is constructed on the basis of relative distribution. This is done thusly. For every threshold its corresponding completeness and accuracy are determined and they are plotted as coordinates - completeness on the axis of abscissas, accuracy on the axis of ordinates. The locus of the points whose coordinates have been obtained thusly is the graph of Smart for the given distribution.

It is possible to show that on the graph of Smart the initial distribution is restored uniquely, if it was Smart and to within similitude, if it was Taub. Therefore the graph of Smart can serve as the evaluation-description.

The graph of Smart can also serve as the evaluation-scale, but only for Smart systems. Specifically, it is possible to consider that for complete Smart systems the distribution is better, its the higher graph. In fact, it is possible to show that if the graph of one Smart distribution is higher everywhere than the graph of another, then at any fixed threshold both completeness and accuracy in the first distribution will be higher than in the second.

It can be done in the following manner. Assume with a certain inquiry and array of documents the graph of distribution D5 lies everywhere higher than the graph of distribution D6. Let us select a certain threshold. At this threshold distribution D6 will give a certain completeness and accuracy. At this same threshold distribution D5 cannot give either equal or less completeness: the first is evident, let us show the second. In fact assume at the selected threshold D5 gives less completeness. This means that with the same quantity of issued documents (thresholds equal!) D5 gives less than necessary. Then in distribution D5 let us move the threshold to the right so that we obtain the same completeness as also in D6 at the initial threshold. In this case in D5 the total amount of issued documents will become greater than in D6 at the initial threshold, and quantity of necessary among them - equal. This means that with the same completeness D5 gives less accuracy than D6, which contradicts the condition about the location of graphs. Therefore at initial threshold the completeness in D5 could be only higher than in D6, and this - because of the equality of common distribution - will attract greater accuracy.

Thus the graphs of Smart utilized in the indicated manner as the evaluation-scale for individual distributions in complete Smart systems, are sensible for all those points of view, for which evaluations, one way or another based on completeness and accuracy, are sensible.[1]

At the same time the analogous application of the graphs of Smart

[1] The graphs of Smart can also be constructed for Taub systems. However, it is unclear what it would be possible to use in this case as the evaluation-scale: is it possible to indicate such Taub distributions, of which one at any threshold gives greater completeness and accuracy while its graph lies everywhere lower than the graph of the other distribution.

as an evaluation-scale for distributions in incomplete Smart systems is hazardous, since it is not difficult to visualize such a situation in which Smart graphs prove to be a nonsensible evaluation of incomplete systems. Assume for example during work with an incomplete system we cease the examination upon achievement of a preassigned value of accuracy and evaluate the result for completeness. Let us consider the following two distributions D7 and D8.

+ − + + − − − − ... (Д7)
− − − + + + − − ... (Д8)

The following graphs in Fig. 2 correspond to these distributions. If now the survey is conducted up to the achievement of an accuracy of 1/2, then in distribution D7 we will obtain completeness 1/3, and in distribution D8 - completeness 1. Thus from the expounded meaningful point of view D8 is better than D7, although the graph of D7 everywhere is higher than the graph of D8.

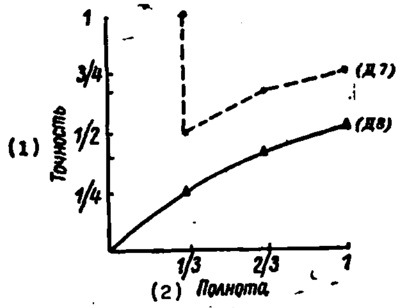


Fig. 2.
KEY: (1) Accuracy; (2) Completeness.

Let us note that as far as is known to us, Smart himself does not conduct such an analysis of the applicability of his graphs as an evaluation-scale.

For the description and comparison of retrieval systems Smart introduces averaged graphs. Such a graph is constructed from the distributions obtained with the help of the evaluated system for a certain array of inquiries on the fixed array of documents. It is done thusly. On the axis of abscissas any values of completeness (Smart uses values 0.1; 0.2, etc.) are plotted. For each of these values in every distribution a threshold (apparently, minimum) is determined which ensures the given value of completeness, and the value of accuracy corresponding to this threshold. For every value of completeness obtained in this manner the values of accuracy are averaged for all distributions which participate in the construction of the graph, and the averaged values of accuracy are plotted on the graph as ordinates.

If we look at the averaged graph obtained as at the evaluation-description, then it is evident first of all that it is not suited (it is not effective) for the description of complete systems. In fact, if we guaranteed by that or other means in every distribution, i.e., for every inquiry, one and the same completeness, then the graph of Smart would show us what would be the average accuracy for all inquiries in this case. But in order to obtain one and the same completeness in different distributions, we must for every distribution, i.e., for every inquiry, establish its threshold, which is unreal because under contemporary conditions a complete system can work only with a threshold which is common for all inquiries. Therefore the natural means of construction of an averaged graph for complete systems is the averaging of completeness and accuracy for all inquiries at all possible fixed thresholds. The values of completeness and accuracy, obtained for every threshold and averaged for all distributions, are used the coordinates of the points of the averaged graph. Such a graph would give, for example, the answer to the question of whether it is possible for the given complete system to select a threshold such that the average values of completeness and accuracy would fall in the assigned limits.

If now we consider the question of the utilization of the averaged charts of Smart for the description of incomplete systems, then also in this case the naturalness of the procedure of averaging proposed by Smart is doubtful. In fact, during work in an incomplete mode the moment of the curtailment of survey - the analog of the threshold - is established by man and it can change from inquiry to inquiry. In this case man should (by definition of an incomplete system) consider the relevancy of the documents being looked over. But if this relevancy was not used, then he cannot use it for the determination of completeness, since he does not know the total number of necessary documents. An effective description of incomplete systems could be the graphs, obtained by the averaging of completeness in the case of fixed accuracy. Such a graph could give the answer to the question on which average completeness we can calculate if we cease the survey on the assigned accuracy. Therefore, it is possible to visualize the reasonable points of view, at which such graphs would give useful information about a system. The Smart method of averaging in the case of fixed completeness seemingly does not give the possibility to effectively use the information about the retrieval system contained in the graphs from any reasonable point of view. Thus the impression is created that Smart graphs (in the Smart method of averaging) are not an effective evaluation-description either of complete or incomplete systems. Unfortunately Smart himself does not give any indications which could help during attempts to use his graphs as evaluation-descriptions.

Let us now move on to the question of the utilization of averaged Smart graphs as an evaluation-scale, i.e., for the comparative evaluation of various retrieval systems.

Smart uses his graphs extensively for this purpose considering it self-evident that if the graph of the first system lies wholly higher than graph of the second, then the first system obtains a higher value of evaluation. When averaged charts intersect, Smart considers it possible to equalize the corresponding nonintersecting sections of the graphs and in this case to speak about the superiority of one of the two compared systems (which these did not mean) "in the given area" - for instance, in the area of high or low completeness.

In accordance with what was said earlier, the examination of the particular question is reduced to the investigation of the area of soundness of Smart graphs. Since the Smart graph for a system can be obtained by means of certain averaging directly from individual distributions (and not by means of the averaging the evaluations of individual distributions), then the question in turn can be formulated as the question of the sensibleness of this or that means of averaging.

Let us consider the soundness of Smart graphs relative to Smart systems. Imagine that we have two absolute complete systems. Let us name these systems C1 and C2. Further let us imagine that these systems are tested on a certain control array of inquiries, whereupon C1 orders the documents by absolute ranks equally for all inquiries - in the first place stands one unnecessary document, behind which follow first all the necessary ones, and then the remaining unnecessary ones. Let us assume that here all the necessary documents fall in the segment from the 1st to the k-th rank, and unnecessary (besides the first) - from (k + 1) to (k + m) rank. Further assume system C2 orders the documents differently on two halves of the array of inquiries: on the first half the documents are ordered just as in system C1, excluding only that the unnecessary document standing in first place changes places with the last from the necessary; on the second half of inquiries the documents are arranged just as on the first half, but with the common shift to the right by k ranks.

Now it is clear that from the point of view of "patent purity" described above system C2 is worse than C1 because the minimum threshold which ensures one hundred percent completeness for system C2 should be assigned in the region of rank 2k, which will draw in great noise on the first half of the inquiries.

Smart graphs for systems C1 and C2 appear in the following manner (Fig. 3):

Hence it follows that from the point of view of "patent purity" the graphs of Smart are not a sensible evaluation for absolute complete systems.

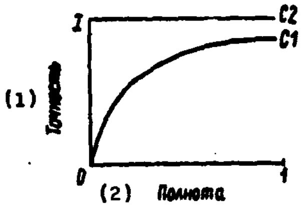


Fig. 3.
KEY: (1) Accuracy; (2) Completeness.

Here it is appropriate to make a short digression. Smart works with relative distributions, without posing the important, generally speaking, question of whether this concept can be the adequate facility of research in general. This question is connected with another: does the case occur, when transition from absolute distributions to relative gives rise to loss of substantial information.

For a clarification of this it is instructive to consider systems C1 and C2 as relative. In this case the shift on the scale of absolute ranks disappears, and system C2 proves to be ideal, and C1 somewhat worse because of the first unnecessary document. This shows that we can lose very substantial information about an absolute system if we are speaking about it in terms of relative distributions, for example, in terms of Smart graphs. Let us note that the importance of this loss is detected only when we have shifted from the evaluation of individual distributions to the evaluation of systems: in the individual distribution for any absolute threshold it is always possible to indicate its equivalent relative threshold, which, generally speaking, cannot be done for a set of averaged distributions of the system.

Let us return to the fundamental presentation and let us consider now the graphs of Smart relative to complete relative systems.

Assume Smart systems C3 and C4 are tested on two control inquiries in a complete relative mode. Assume systems C3 and C4 gave respectively the following distributions of documents on the relative scale:

```
— — — — — — + + + — —  1 запрос (1) }
+ + + + + + — — — — —  2 запрос (1) } Система C3(2)
+ + + — — — — — — — —  1 запрос (1) }
— — — + + + + + + — —  2 запрос (1) } Система C4 (2)
```

KEY: (1) inquiry; (2) System.

If we construct the completeness - accuracy graph, averaging the completeness and accuracy at fixed relative thresholds from 1 up to 11, as it seems natural to us to do for complete systems, then for both systems we will obtain the same graph (Fig. 5).

If we construct the graphs by averaging according to Smart, then the graphs which result will be such (Fig. 4):

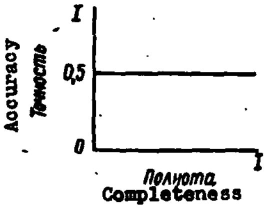


Fig. 4.

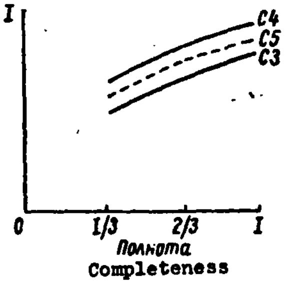


Fig. 5.

This example, as also the previous one, shows that the method of Smart is not the only possible one and that other methods exist for the construction of averaged graphs which give different results than Smart in the comparison of systems. Therefore the question of the areas of soundness of various methods of averaging naturally arises.

It can be discussed thusly. For any complete system - by definition - a certain threshold is fixed which separates distribution from nondistribution. Since this threshold should be determined not in the terms of relevancy, but for relative systems, it seems it can be assigned only in the terms of the number of documents or ranks distributed (which in the case of Smart systems is one and the same). It is natural to consider that if from two compared systems one at any threshold it gives the best average completeness with no worse average accuracy or the best average accuracy with no worse average completeness, then it is better than the second. This actually means that it is natural to average out completeness and accuracy at fixed thresholds. From this viewpoint the first graph of the last example in question is natural, and the systems compared identical. With respect to this the second graph, obtained by Smart averaging, is abnormal, and the comparative evaluation of systems C3 and C4 given by it is distorted.

In other words, from the described point of view averaging in the case of fixed thresholds is sensible, and Smart averaging - senseless.

However, this point of view is not the only one possible, and it can be discussed in another way. Assume for example that the user requires that on every inquiry he obtains at least two necessary documents and as little distribution as possible. For such a user system C4 (as complete) is better than system C3 (also as complete), which corresponds to the evaluation from the graphs of Smart and contradicts the evaluation obtained by averaging on fixed thresholds. All this once more illustrates the position already expressed by us that the criterion of the soundness of one or another evaluation can be only the agreement of the scale of compared objects given by it with the meaningful accepted scale. This latter can be developed only in resolving any concrete problem, and therefore the soundness or senselessness of one or another evaluation becomes dependent on that problem which determines the conditions of application of the evaluation in question.

The impression can be put together that the above described point of view of the user, who wishes to obtain as little distribution as possible and in this case to obtain for every inquiry no less than two necessary documents, enters into the area of soundness of Smart averaging. This, however, is not so.

In order to show this, let us perform the following. Let us take system C5, and assume on the same control inquiries on which systems C3 and C4 were examined it gives the following distributions:

```
+ + + − − − − − − − − 1 запрос (1)
                                   система C5 (2)
− − − − − + + + + + + 2 запрос (1)                  KEY:  (1) inquiry; (2) system
```

From the point of view of the user in question system C5 is worse than system C3 because it gives in the case of no less than two necessary documents for each inquiry a total of 9 unnecessary against 8 unnecessary for system C3. If we construct a Smart graph of system C5, then it lies everywhere higher than the graph of system C3 (see Fig. 5).

The examples given in the examinations conducted above can be considered artificial, and therefore unconclusive. But then the question arises, generally which examples to consider as conclusive, or even - the common question of the principles of confirmation or disproof (substantiation or discrediting) of values. It is possible to relate differently to this to problem, but our examination shows, in any case, that it cannot be disregarded.

Smart in essence disregards this problem. At our disposal there are two excerpts from the publications of Smart which contain something similar to a substantiation of the completeness - accuracy graphs used by him. In the first of them he speaks about graphs as about an evaluation-scale, in the second - as about evaluation distribution. Here they are:

In the first excerpt Smart writes: "in order to give some indications of the quality of the systems which would yield readily to a comparison with previously published data, we compute the standard completeness and accuracy ..."

Further Smart gives the description of his completeness - accuracy graphs and then continues:

"these graphs in such an accurate form have been introduced by Cleverdon... The procedure of averaging described above is distinguished from Cleverdon"... (Smart describes precisely in what and continues): ... "Although the actual computation is conducted thus from several various points of view, the graphs given by us should nevertheless yield to comparison with the published Crenfield material" ([1], p. 219).

The differences in the method of construction of Crenfield and Smart graphs are not limited to those which Smart indicated, but independent even of this they are sufficiently significant that the question of the possibility of the direct comparison of graphs would deserve special substantiation.

The second excerpt is such: "such graphs effectively neutralize the evaluations of retrieval methods based on various inquiries and they can be used with benefit for the selection of those retrieval methods which approach certain specific conditions of functioning. Thus, for instance, if it is required to select a procedure which would give maximum completeness, then one ought to examine the graphs while being limited to the area of high completeness; analogously if it is required to obtain only necessary documents, then the area of high accuracy is important" ([2] p. 12). Such a substantiation is evidently insufficient. We will be limited in this connection to reference to our examples of C3-C5, which remain valid even in the case of the requirement of full completeness; there is no doubt that also for the case of complete accuracy it would be possible to devise the appropriate examples.

In conclusion let us make the following observation.

Up to now all the evaluations of retrieval systems examined by us were obtained in one way or another on the basis of fixed arrays of inquiries and documents. It is completely evident that the evaluation of **retrieval system** cannot be spoken of without investigating the question of the presence of the control array of documents and inquiries, i.e., the stability of the value of the evaluation relative to the change both of the array of inquiries and the array of documents. With the entire importance and urgency of this question it emerges beyond the frames of the present examination.

§ 4.

Pollock is one of a few authors known to us who speak sufficiently definitely about the communication between formal evaluation and the meaningful point of view. The most characteristic in this respect is the following excerpt:

"Maybe the most direct method of approach to this problem (the discussion concerns which properties should be possessed by a good evaluation - V. Ch. and D. L.) is to explain, why in the final analysis distributions are necessary - what it is planned to do with them. Only having answered this question is it possible to explain the importance of the order of documents inside the distribution or the importance of the quantity of the elements of distribution, etc...." ([3], p. 390).

Pollock subsequently gives three examples of that which he considers as possible points of view (one ought to keep in mind that below we are describing the content of the work of Pollock in our terms and therefore let us bear the responsibility for the correctness of interpretation).

The first of them Pollock characterizes by the fact that the user does not wish to look over more than three documents, but he wishes that among them there would be at least one necessary one; this point of view is not invariant relative to the similarity transformation.

The second is characterized by the fact that the user is prepared to look over many documents - Pollock gives the number 400, even if among them there will be comparatively few necessary ones - Pollock gives in this connection the number 100. It remains unclear if the second point of view differs from the first only by the fact that in it considerably larger numbers are figured or by the fact that it is invariant relative to the similarity transformation.

The third point of view is distinguished by the fact that the user wishes to use a system in an incomplete mode and to look over distribution until based on that or other considerations he is content with the information obtained. Further Pollock clarifies that from this viewpoint the distribution is good in the case when the documents are arranged in the order of decreasing relevancy. This point of view is invariant relative to the similarity transformation.

Further Pollock expresses the thought that the evaluation should consider the diversity of possible points of view.

It is very significant to us that Pollock realized this thought: he introduces into the evaluation proposed by him in the same work the certain parameter which in practice proves to be simply relative threshold (in the distribution described below L').

We already stated that the idea of parametric representation of the points of view is interesting and urgent to us, however, it is more than doubtful that such a representation can be constructed in a natural way as one-parameter and already it seems entirely obvious that relative threshold cannot be such a parameter. Last it is evident at least from that that the difference between the points of view, invariant relative to the similarity transformation, and those of them which relative to such a conversion are noninvariant, cannot be reflected by relative threshold.

Pollock constructs his evaluation in the form of the relation

$$\mu(n) = \frac{f(n)}{f^{*}(n)},$$

which is determined for all values of n from 1 to N, where N is the number of documents in the distribution. Primarily from here it is clear that the evaluation of Pollock is parametric in the same sense as also the Smart completeness-accuracy graphs which possibly should be considered as the first attempt to construct parametric evaluations.

Pollock functions f and f* can be determined in the following manner.

Assume there is an array of documents and any inquiry. It is assumed that for every document D_i(i = 1.2..., N, where N is the number of documents in the array) weight v_i has been assigned. It characterizes the relevancy of this document relative to the fixed inquiry. Further let the investigated retrieval system have a given array of documents on the given inquiry in the distribution L, for which it is necessary to construct a Pollock evaluation. Distribution L can be both Smart and Taub. Now we will construct two auxiliary distributions L' and L*.

Distribution L' is an "ideal" distribution of our array of documents on the given inquiry. In this distribution the documents are arranged by decreasing weights v_i, and documents with equal weights have been ordered between themselves arbitrarily.

Distribution L' represents a Smart distribution which is obtained from L by means of the arbitrary ordering of documents within the limits of every rank of distribution L. It is clear that if L is a Smart distribution, then L' coincides with L. With distribution L' we connect the auxiliary system of weights v_i', which is constructed in the following manner: for every rank of distribution L the arithmetic mean of the weights of the documents entering this rank is calculated, and this arithmetic mean is considered the auxiliary weight of every document entering this rank. Therefore all documents entering one rank of distribution L have the same auxiliary weight, and the sum of their weights is equal to the sum of their auxiliary weights.

If now in distribution L' we sum up the auxiliary weights of all documents up to rank n inclusively, then we obtain the value f(n); if we do the same in distribution L* with the weights, then we obtain the value f*(n).

The sense of evaluation μ(n) is clarified most simply for the case when distribution L is Smart. Then L' coincides with L and the auxiliary weights coincide with the weights. Under this assumption f(n) reaches a maximum, equal to f*(n), if L coincides with L*. In this case μ is identically equal to 1, which is a maximum for it. Any change in the location of documents as compared with distribution L* leads to the fact that for any values n f(n) becomes less than f*(n), and therefore μ(n) becomes less than a unit.

Pollock does not say how his evaluation of μ should be used for the comparison of distributions, however it is possible to assume that he considers distribution L to be better, the higher its graph μ(n) lies. In this connection let us consider the distributions

—+——— (Д9)

and

—++—— (Д10)

(+ has the weight of 1, - has the weight of 0).

Graphs μ(n) for these distributions take the following form (Fig. 6). We see in such a way that the graph of distribution D9 lies higher than the graph of distribution D10, and therefore distribution D9 receives a higher formal evaluation than distribution D10. From a meaningful point of view the advantages of distribution D9 at least are not obvious: it is much simpler to visualize situations when distribution D10 is meaningfully better than such when it is worse than the distribution D9. For this it is sufficient, for example, in a complete mode to place the threshold equal to three and to compare the noise, or in an incomplete mode to suppose that survey is ceased at two subrows of operating unnecessary documents, and noise is also compared.

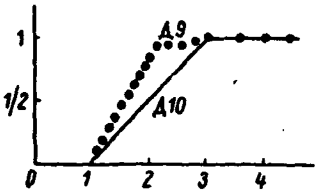


Fig. 6.

Not less significant things are detected, if we compare distribution D9 with distribution D11:

——++————————(Д11)

It is possible to consider that distribution D11 is obtained from distribution D9 in the case of an invariable retrieval system and the same inquiry by means of the trivial doubling of the array of documents. Meanwhile the graph of distribution of D11 takes the form (Fig. 7). Consequently there where the graph of distribution D9 has been determined, it is higher than the graph of distribution D11, and the evaluation of Pollock proves to be noninvariant relative to the similarity transformation and its partial case - trivial doubling. From this it further follows that from the three points of view given by Pollock as an example one - the third - does not knowingly enter the area of soundness of its formal evaluation.

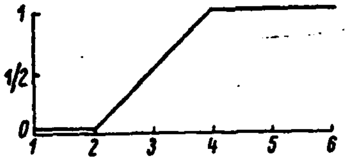


Fig. 7.

But the matter is not limited to this. If we turn to the first of the points of view expressed by Pollock then it is not difficult to notice that it in general does not determine any meaningful evaluation of retrieval systems because it relates not to retrieval systems, but to retrieval services. In fact, let us assume that the user, standing on the first point of view, uses the system which on the theme which interests him gives a systematic error. Let us assume that on a small array of documents this gives rise to distributions of the type D9, and the user remains satisfied. If now in the course of time the array documents is supplemented, but preserving its thematic structure, then the distributions of the type D9 should gradually be replaced by distributions of the type D11, and user will change the meaningful evaluation from satisfactory to unsatisfactory in spite of the fact that system remained the same. Hence it is clear that, relying on the first of the Pollock points of view, the user evaluates not the retrieval system itself, but the service which includes the retrieval system and the array of documents.

This circumstance by itself in no way diminishes the point of view in question, since retrieval services should be the object of comparison and evaluation the same as the system. But if one considers that Pollock introduces his formal evaluation **precisely as the evaluation of systems**, then it turns out that also the first of the points of view cited by him should not enter the area of soundness of its evaluation. The fact that Pollock evaluation (noninvariant, as we showed above relative to the similarity transformation) in general cannot serve as an evaluation for systems in this case does not save the positions.

The second point of view is formulated by Pollock insufficiently clearly, so its examination is connected with the known risk of misunderstandings. The most probable are the following two alternative interpretations of it. In the first place it is possible to consider that this point of view is a variety of the first, and therefore relates not to systems, but to services. In the second place it is possible to consider that it consists of the fact that the user does not limit the dimensions of distribution, but requires that the fraction of necessary documents in the distribution would be no lower than a certain specific number; in this case the second point of view proves to be invariant relative to the similarity transformation. Thus a great probability exists that also the second point of view does not enter the area of soundness of Pollock evaluation.

This strange situation is the result of the fact that Pollock in reality does not concern himself completely with the question of the conditions of applicability of his evaluation, and the connection of the evaluation with the point of view proclaimed by him remains unrealized in his work.

In conclusion let us briefly note that Pollock, having constructed his evaluation for distributions, says nothing about how to construct an evaluation for systems. Meanwhile, as it was shown above, the area of soundness and effectiveness of the evaluation for systems substantially depends on the method of averaging.

§ 5.

The absence of a serious examination of the questions connected with the high quality of evaluations gives rise to the fact that various evaluations are introduced and the laborious computations of the values of these evaluations are conducted in a known sense for nothing. The demonstrative expression of this lamentable state of affairs is the healthy intuitive distrust, with which the practical workers in the field of creation and application of information retrieval systems relate to all possible formal evaluations, especially in those cases when the results of their application diverge from intuition. A completely indicative example can be the derivations from the 2nd Crenfield experiment (see [6]). In this vast and in many re respects remarkable experiment the retrieval effectiveness of 33 various search systems was compared on the basis of utilization of an evaluation-scale, reminiscent of the noralized completeness R_N of Smart. The search systems compared differed by their languages, which were varied from the simplest single-term to languages with complex bases and textual relations. As a result of the experiment the highest evaluation was received by a language, the single base relation in which was the equivalency of key words having a common root (root); the single-term language without any base or textual relations turned out in third place (from 33 languages compared!) Whereupon the value of evaluation for it was only insignificantly inferior to the value of evaluation for the language acknowledged as best (respectively 65.82 and 65.00). However, thus far, as far as it is possible to judge, no one yet intends to declare on this foundation the single-term language the best (or at least almost the best) from the possible retrieval languages.

Insufficient acquaintance with the materials of the 2nd Crenfield experiment does not permit us to compose a precise representation about, precisely on which measure the evaluation used is responsible for such an unexpected result and in which measure its field of soundness limits its generality. However, this example, as it seems to us, visually illustrates the thesis which we strove to justify in the present work: to introduce an evaluation for retrieval systems is easy; it is much more difficult, but no less it is necessary, to persuasively show its high quality and to indicate the conditions of its effective application.

The aforesaid does not mean, it goes without saying, that all the experiments in evaluations conducted up to now, especially the comparative evaluations of retrieval languages and systems (Smart, Cleverdon, Sokolov [7]), are senseless or useless. On the contrary, only these experiments which signify a new stage in the development of information theory, made it possible to reveal the deficiencies in the presently available apparatuses of evaluation and comparison. However, the further utilization of the present apparatuses without a fixed study of their properties and limitations is inadmissible for us.

In this article for the analysis of the formal values of retrieval systems we used the concept of soundness and effectivenesses which are based on the concept of meaningful evaluation or point of view. We consider as one of the most important trends of further research in the field in question the creation of a language for the description of the points of view and, more generally, retrieval situations, and also the study on the methods for the precise determination of the fields of soundness and effectiveness of the evaluation of retrieval systems.

In our examination we used two methods for the study of evaluations: the disproving examples and invariance relative to conversions. The conclusive force of these methods, generally speaking, can be placed under doubt. Thus, disproving examples, even very clear by themselves, can hardly compromise an evaluation, if under the actual conditions of its utilization they are a rare exception. Further it is possible that two evaluations, invariant relative to various conversion groups, under any actual conditions will be sufficiently close in that sense, that for the same systems they will give the same - or almost some - values. For example, Cleverdon considers it possible to compare the results of his experiments with the results of the experiments of Smart, apparently being based on the intuitively perceptible resemblance of evaluations used by them, although these evaluations - as can be shown - are invariant relative to the different conversion groups. Is such a form of actions justified, are the counterexamples given by us above conclusive - this can be solved only as a result of special investigations of the conditions of the actual application of the values of retrieval systems, the specific value of special cases, the concept of the resemblance of evaluations, etc. If our work serves as a stimulus for such studies, we will consider our mission fulfilled.

BIBLIOGRAPHY

1. Salton G. The Evaluation of Automatic Retrieval Procedures-Selected Test Results Using the SMART System. Amer. Decum., July 1965, 16, № 3, 209—222.
2. Salton G. and Lesk M. E. Computer Evaluation of Indexing and Text Processing. «J. Assoc. Computing Machin.», Jan. 1968, 15, № 1, 8—36.
3. Pollock S.-M. Measures for the Comparison of Information Retrieval Systems. «Amer. Docum.», Oct. 1968, 19, № 4, 387—397.
4. Бернштейн Э. С., Лахути Д. Г., Чернявский В. С. Некоторые вопросы построения дескрипторных информационно-поисковых систем. «НТИ», 1963, № 1, 31—33.
5. Чернявский В. С., Лахути Д. Г., Бернштейн Э. С. Информационно-поисковые системы. Сборник лекций: «Теория и практика научно-технической информации», М., 1969, 220—313.
6. Cleverdon C. The Crenfield Tests of Index Janguage Devices. «Aslib Proc.», 1967, 19, № 6, 173—194.
7. Соколов А. В. Исследование потерь информации и информационного шума в дескрипторных информационно-поисковых системах. «НТИ», 1965, № 12, 23—28.
8. Cooper W. S. Expected Search Length: a Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems. «Amer. Docum.», Jan. 1968, 19, № 1, 30—42.

Article received by the editors
27 June 1969